

State-Dependent Structural Shocks and Temporal Generalization in Corporate Default Prediction

Li Fan *

School of Computing and Data Science, Xiamen University Malaysia, Sepang, 43900, Malaysia

* Corresponding author: (Email: DSC2409010@xmu.edu.my)

Abstract: This study examines whether financially structured data augmentation can enhance corporate default prediction under strict temporal generalization. Using an anchored temporal split—1999–2004 for training, 2005–2006 for validation, 2007–2009 for stress testing, and 2010–2018 for post-stress evaluation—it first documents a limitation of current practice: static augmentation fails to improve out-of-time discriminative power. The raw baseline attains a future mean AUC of 0.6858, whereas SMOTE and static constrained augmentation yield only 0.6790 and 0.6798. The paper then introduces a state-dependent structural shock framework that injects economically disciplined deterioration paths into stable-regime training data. The calibrated acute/persistent design achieves the highest future mean AUC, 0.6947, above the regime-feature baseline of 0.6927. Boot-strap evidence supports this AUC gain directionally rather than uniformly, with stronger support in the post-stress window than in the stress window. By contrast, the gain does not extend to future mean PR-AUC. Decision simulation further shows that the practical advantage of shocks is strongest under conservative screening rules rather than aggressive recall. Dynamic stress rehearsal therefore improves out-of-time ranking selectively, but not universally across metrics, baselines, or operating points.

Keywords: Corporate default prediction; Anchored temporal split; Out-of-time robustness; Distribution shift; State-dependent structural shocks.

1. Introduction

Corporate distress prediction remains important because it links accounting information to credit allocation and risk management. Classical ratio- and logit-based studies established that financial statements contain substantial default information [1, 2]. Later work shows that machine-learning classifiers can improve predictive performance, especially in nonlinear settings [3]. Yet two methodological challenges remain fundamental. First, bankruptcy is a rare event. Second, the mapping from financial features to distress is state-dependent.

Most augmentation methods mainly address the first problem. SMOTE and related procedures enlarge the minority class within the observed feature space [4, 5]. What remains unclear is whether such gains survive when the deployment environment differs from the training regime. This concern is especially salient in credit risk, where macroeconomic conditions shape both leverage and default dynamics [6–8]. More broadly, concept drift and underspecification caution against treating in-distribution validation as sufficient evidence of real-world robustness [9, 10]. In financial machine learning, the limitations of conventional cross-validation under temporal dependence and leakage are now widely recognized, which further motivates an anchored temporal design [11].

Relative to the recent literature on constrained tabular data generation, which primarily emphasizes constraint compliance and synthetic-data utility, the present study occupies a narrower yet practically important niche: it asks whether financially structured augmentation improves anchored temporal generalization in corporate default prediction [12–14]. Preliminary experiments under conventional random splits suggested that static constrained augmentation could deliver modest in-sample gains. Once evaluation was aligned with the forward-looking nature of credit prediction

through a strict anchored temporal split, however, that apparent advantage largely disappeared.

The contribution of this study is threefold. First, it shows that anchored temporal evaluation overturns the apparent benefit of static augmentation. Second, it proposes a state-dependent shock design that encodes regime-dependent deterioration logic rather than merely filtering financially plausible synthetic points. Third, it shows that the resulting gains are selective: the evidence is stronger relative to raw, SMOTE, and static constrained baselines than relative to a regime-feature baseline, and the practical value is most visible under conservative screening rules.

2. Data

The historical base sample is a pre-assembled U.S. corporate distress panel for 1999–2018. It contains 78,682 firm-year observations from 8,971 firms and a binary status label with values alive and failed. Because the source data do not distinguish Chapter 11 filings, technical defaults, or other distress subtypes, the positive class is interpreted strictly as dataset-defined corporate failure rather than as a hand-audited legal bankruptcy event. The panel is unbalanced and observed at the firm-year level.

A second SEC-based reference sample for 2019–2025 is used only to validate cross-source feature harmonization. While anonymized firm identifiers preclude linking to external databases such as CRSP or Compustat, this constraint also isolates the methodological properties of temporal robustness from unobserved firm-specific metadata. As shown in Table 1, the anchored temporal design separates the training, validation, stress, and post-stress windows in a strictly forward-looking manner.

Table 1. Anchored temporal design

Window	Years	Firm-years	Firms	Positives	Positive rate
Train	1999–2004	28,847	6,426	2,431	0.0843
Validation	2005–2006	8,333	4,477	745	0.0894
Stress test	2007–2009	11,609	4,370	854	0.0736
Post-stress test	2010–2018	29,893	5,192	1,190	0.0398

The analysis uses a unified five-feature representation. Because the historical source provides anonymous fields, the variables should be interpreted as economically motivated proxies aligned to SEC ratio definitions. As shown in Table 2, the empirical design is built on a unified five-feature representation spanning liquidity, cash flow, profitability, and capital structure.

Table 2. Five-feature representation

Feature	Historical SEC-style anchor field		Interpretation
Cash ratio	X8 Cash / Liabilities		liquid cash buffer
OCF to assets	X12	Operating cash flow / Assets	internal cash generation
ROA proxy	X6 Net income / Assets		profitability
Operating CF ratio	X4	Operating cash flow / Liabilities	liability coverage by cash flow
Equity ratio	X13 Equity / Assets		balance-sheet resilience

The retained features collectively span liquidity, operating cash generation, profitability, and capital-structure resilience. Historical raw features are winsorized at the 1st and 99th percentiles and then standardized using parameters estimated on the training sample only.

3. Method

Let X denote the five-dimensional feature vector and $Y \in \{0, 1\}$ the distress label. The empirical premise is that the conditional geometry of distress is not regime-invariant:

$$P_{stable}(X | Y) \neq P_{stress}(X | Y) \quad (1)$$

Equation (1) is not posited as a formal structural mechanism; rather, it serves as an empirical heuristic indicating that classifiers estimated on stable-period patterns may be misaligned with crisis-period dynamics.

Five training variants are compared: raw baseline, SMOTE, static constrained augmentation, regime-feature baseline, and state-dependent structural shocks. Static constrained augmentation first generates minority-class candidate points through SMOTE-style interpolation and then filters those proposals through four static rule groups. To isolate the effect of feature-space geometry, the regime-feature baseline adds a binary regime indicator to the five financial features; the indicator is set to 0 in the train and validation windows and to 1 in the future evaluation windows, so that it functions as a coarse proxy for a non-training regime rather than as a probabilistic state model.

The number of additional minority cases is set to

$$n_{aug} = \min\{n_0 - n_1, n_1\} \quad (2)$$

Where n_1 and n_0 denote the minority and majority counts in the train window. In 1999–2004, $n_1 = 2,431$, so the augmentation target is 2,431 synthetic minority observations.

The shock design is explicitly state-dependent and economically disciplined. For feature c and regime $r \in \{\text{acute}, \text{persistent}\}$,

$$\tilde{x}_{i,c}^{(r)} = x_{i,c} - u_i^{(r)} \lambda_c^{(r)} \Delta_c \mathbf{1}(i \in \mathcal{A}) \quad (3)$$

Where $\mathcal{A}$ is the set of proposed minority-class candidates generated from training-window distress observations, Δ_c is the baseline deterioration magnitude for feature c , defined as the winsorized mean difference between bankrupt and alive observations in the training window, $\lambda_c^{(r)}$ is the regime-specific multiplier, and $u_i^{(r)} \sim U[0.8, 1.2]$ introduces mild severity heterogeneity. In the acute regime, larger multipliers are assigned to liquidity and operating-cash-flow features to reflect the cash-squeeze and rollover-risk dynamics that prevail in sudden crises [6, 8]. In the persistent regime, larger multipliers are assigned to profitability and equity-capital features to capture the slower erosion of earnings capacity and balance-sheet resilience [7]. The final synthetic observation is projected back to the admissible region:

$$x_i^{shock} = \Pi_{\Omega}(\tilde{x}_i^{(r)}) \quad (4)$$

Where $\Pi_{\Omega}(\cdot)$ is the clipping/projection operator and Ω is the feasible financial region implied by the maintained constraints.

The static constraint system contains four rule groups: cash-flow/liquidity bounds, profitability bounds, equity-structure bounds, and cross-feature consistency checks. Logistic regression, random forest, and gradient boosting are estimated with fixed ex ante hyperparameters to avoid look-ahead tuning. The principal forward metric is with an analogous definition for future mean PR-AUC.

$$FutureMeanAUC(m) = \frac{1}{2}(AUC_{m, stress} + AUC_{m, post}) \quad (5)$$

To assess whether mean improvements are stable rather than purely descriptive, the paper performs bootstrap-based pairwise comparisons for the shocks design against raw baseline and regime-feature baseline. The evaluated metrics are AUC and PR-AUC, and the evaluated future groups are the stress window, the post-stress window, and their mean. The paper also conducts decision-oriented evaluation using top- q screening and threshold-based screening. The corresponding decision metrics are precision, recall, false-positive burden, and cost proxy. Together, these additions clarify whether ranking gains are statistically directional and how they translate into operating-point implications.

4. Results

4.1. Main Anchored Temporal Results

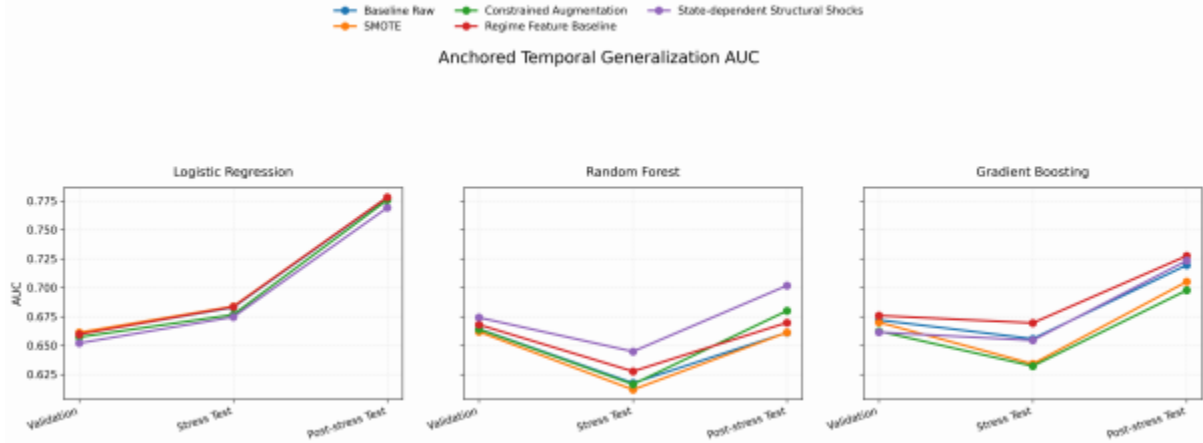
Three patterns emerge from Table 3. First, the raw baseline outperforms both SMOTE and static constrained augmentation on future mean AUC, indicating that static minority-class expansion fails to improve temporal discriminative power. Second, the shocks design attains the highest future mean AUC, 0.6947. Third, the regime-feature baseline achieves the highest future mean PR-AUC, so the shocks design cannot be described as uniformly dominant.

Table 3. Anchored temporal generalization: mean results across model classes

Method	Validation AUC	Stress AUC	Post-stress AUC	Future Mean AUC	Future Mean PR-AUC
Raw baseline	0.6654	0.6521	0.7194	0.6858	0.1233
SMOTE	0.6644	0.6432	0.7148	0.6790	0.1206
Static constrained augmentation	0.6612	0.6418	0.7178	0.6798	0.1153
Regime-feature baseline	0.6679	0.6602	0.7252	0.6927	0.1266
State-dependent structural shocks	0.6626	0.6580	0.7314	0.6947	0.1216

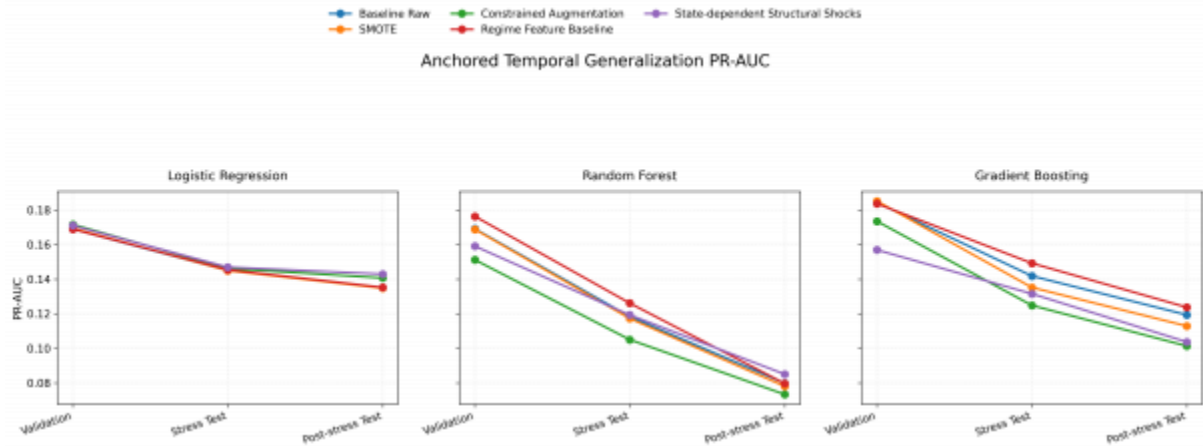
Figure 1 shows the model-level AUC profile. The shocks design attains the strongest aggregate future mean AUC in the mean comparison, but the gain is heterogeneous and does not

extend uniformly across metrics or model classes. The clearest advantage appears in the post-stress period.

**Figure 1.** Anchored temporal AUC by model class

The shock design does not uniformly dominate each split-specific comparison, but it yields the strongest Future Mean AUC once the stress and post-stress windows are considered jointly.

The corresponding limitation appears in Figure 2. The shocks design improves future AUC more clearly than future PR-AUC. In particular, the regime-feature baseline remains marginally stronger on minority-class retrieval.

**Figure 2.** Anchored temporal PR-AUC by model class

The gain of state-dependent shocks is concentrated in ranking quality rather than minority-class retrieval; the regime-feature baseline retains the strongest Future Mean PR-AUC.

4.2. Bootstrap Evidence on Pairwise Differences

Bootstrap evidence reinforces the main result while narrowing the appropriate claim. Relative to the raw baseline, the AUC evidence is directionally positive and more concentrated in the post-stress period. Relative to the regime-feature baseline, the evidence is weaker and narrower. For PR-AUC, the evidence is weaker and mixed. Model-level heterogeneity therefore prevents any universal superiority

claim. As shown in Table 4, bootstrap evidence reinforces the main result while narrowing the appropriate claim.

Table 4. Compact bootstrap summary for the shocks design

Comparison	Stress AUC diff.	Post-stress AUC diff.	Future Mean AUC diff.
Shocks vs Raw baseline	+0.0059	+0.0120	+0.0089
Shocks vs Regime-feature baseline	-0.0022	+0.0063	+0.0020

Relative to the raw baseline, the shocks design delivers a directionally positive Future Mean AUC improvement of +0.0089, with gains of +0.0059 in the stress window and

+0.0120 in the post-stress window. Relative to the regime-feature baseline, the incremental gain narrows to +0.0020, with a small negative difference in the stress window and a larger positive difference in the post-stress window. For PR-AUC, differences re-main mixed and economically modest. The resulting interpretation is therefore selective: the support is stronger for ranking relative to raw and static baselines than for universal dominance over the regime-feature specification.

4.3. Decision-Simulation Evidence

Table 5 makes the business trade-off more explicit. The regime-feature baseline is the most aggressive specification: it delivers the highest recall, but also the highest false-positive rate and predicted-positive rate. By contrast, the shocks design lowers false-positive exposure and yields the lowest cost proxy, but at the expense of lower recall. The reported cost proxy uses the same decision-loss schedule for all methods and is interpreted comparatively.

Table 5. Future-window business metrics (mean across stress and post-stress periods)

Method	Recall	Precision	False-positive rate	False-negative rate	Predicted-positive rate	Cost proxy
Raw baseline	0.4557	0.1996	0.2995	0.5443	0.3078	0.5928
SMOTE	0.4597	0.1695	0.3044	0.5403	0.3126	0.5955
Static constrained augmentation	0.4337	0.2022	0.2800	0.5663	0.2880	0.5882
Regime-feature baseline	0.4929	0.2084	0.3604	0.5071	0.3674	0.6281
State-dependent structural shocks	0.3860	0.1735	0.2241	0.6140	0.2328	0.5620

Decision simulation delivers a parallel conclusion. Under top-q screening, the regime- feature baseline is generally the stronger default when the objective is aggressive recall. Under threshold-based screening, especially in the 0.30–0.50 range, the shocks design more consistently reduces false positives and the cost proxy. This pattern is consistent with the weaker PR-AUC evidence: the shocks design is not the best option for broad minority retrieval, but it is more attractive for conservative screening.

Figure 3 benchmarks the main result against stronger comparators. The left panel shows that adding a simple regime indicator modestly improves anchored performance. The right panel shows that expanding-window re-training sharply improves mean test AUC as later years are incorporated into the training sample. This benchmark is economically relevant, but it answers a different question from the one studied here: re-training with newer data rather than generalizing forward from an early stable window.

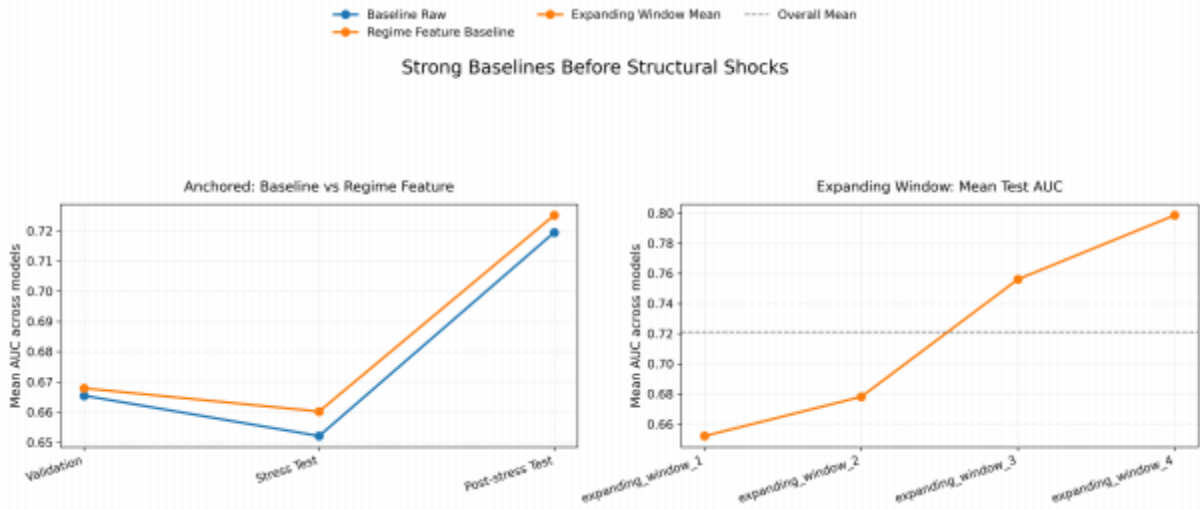


Figure 3. Strong baselines before structural shocks

The left panel shows a modest anchored gain from adding a regime feature. The right panel shows that expanding-window re-training sharply improves mean test AUC as later data are incorporated into the training sample.

Taken together, the evidence suggests a compact interpretation. The method helps more with out-of-time ranking than with minority retrieval. The gain is more visible in the post-stress period. It does not uniformly beat the regime-feature baseline. Its practical value lies in conservative screening rather than aggressive recall.

5. Discussion and Conclusion

Several limitations warrant acknowledgment. The historical source uses anonymized features and identifiers, which precludes direct linkage to external databases and detailed legal-event taxonomy. The auxiliary SEC sample is

unlabeled and therefore cannot serve as a full post-2018 test set. These limitations clarify the scope of the contribution: the study is best interpreted as a methodological investigation of temporal robustness rather than as a production-ready bankruptcy model.

Static augmentation is best understood as improving within-regime plausibility without encoding regime-dependent boundary shifts. That is precisely why its apparent benefit weakens once evaluation moves to a strict anchored temporal design. By contrast, state-dependent shocks provide selective robustness gains in ranking. The bootstrap evidence is more supportive for AUC than for PR-AUC, and the comparison with the regime-feature baseline is more modest than the comparison with raw and static baselines. The comparison is therefore better understood as a trade-off across deployment objectives than as a search for a single

universally superior specification.

The core empirical conclusion is twofold. First, static augmentation does not improve future temporal generalization under anchored evaluation, whereas state-dependent shocks provide selective AUC gains that are more concentrated in post-stress horizons. Second, the decision evidence implies different deployment roles: the regime-feature base-line is more suitable for aggressive recall, whereas the shocks design is more suitable for conservative screening and lower false-positive burden. The contribution of state-dependent stress rehearsal is therefore selective and deployment-dependent rather than universally dominant.

References

- [1] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), 589–609.
- [2] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109–131.
- [3] Barboza, F., Kimura, H., & Altman, E. I. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417.
- [4] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- [5] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- [6] Hackbarth, D., Miao, J., & Morellec, E. (2006). Capital structure, credit risk, and macroeconomic conditions. *Journal of Financial Economics*, 82(3), 519–550.
- [7] Figlewski, S., Frydman, H., & Liang, W. (2012). Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Financial Analysis*, 21, 87–105.
- [8] Bernanke, B. S., Gertler, M., & Gilchrist, S. (1999). The financial accelerator in a quantitative business cycle framework. In J. B. Taylor & M. Woodford (Eds.), *Handbook of Macroeconomics* (Vol. 1, pp. 1341–1393).
- [9] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2022). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 34(10), 4594–4624.
- [10] D’Amour, A., Heller, K., Moldovan, D., et al. (2022). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226), 1–61.
- [11] López de Prado, M. (2018). *Advances in financial machine learning*. Wiley.
- [12] Stoian, M. C., & Giunchiglia, E. (2025). Beyond the convexity assumption: Realistic tabular data generation under quantifier-free real linear constraints. In *International Conference on Learning Representations*.
- [13] Cardei, M., Munoz, J., Barrera, O., Chandrachud, S., & Saha, P. (2025). Constrained tabular diffusion for finance. In *Proceedings of the 6th ACM International Conference on AI in Finance*.
- [14] Stoian, M. C., Giunchiglia, E., & Lukasiewicz, T. (2026). A survey on deep learning approaches for tabular data generation: Utility, alignment, fidelity, privacy, diversity, and beyond. *Transactions on Machine Learning Research*.

Online Appendix

Table 6. Shock calibration

Feature	Δ_c	$\lambda_c^{(acute)}$	$\lambda_c^{(persistent)}$	Acute shock	Persistent shock
Cash ratio	-72.8250	1.25	0.90	-91.0312	-65.5425
OCF to assets	-7.4200	1.20	0.95	-8.9040	-7.0490
ROA proxy	-13.8470	0.95	1.15	-13.1547	-15.9241
Operating CF ratio	-9.4775	1.15	0.95	-10.8991	-9.0036
Equity ratio	-33.1030	0.90	1.15	-29.7927	-38.0685

Table 7. Static constraint groups

Rule group	Summary
Cash-flow / liquidity	Non-negativity floor and historical percentile bounds for cash ratio, OCF/assets, and operating CF ratio
Profitability	Historical percentile bounds for the ROA proxy
Equity structure	Historical percentile bounds for the equity ratio
Consistency	Excludes jointly “too healthy” synthetic observations and implausible cross-feature combinations

Table 8. Model settings

Model	Setting
Logistic regression	solver=liblinear, max_iter=1000, class_weight=balanced, random_state=42
Random forest	n_estimators=120, max_depth=8, min_samples_leaf=2, class_weight=balanced_subsample, random_state=42
Gradient boosting	n_estimators=80, max_depth=2, subsample=0.8, random_state=42

Supplementary Remarks

In the training window, SMOTE generates 2,431 synthetic minority observations. Static constrained augmentation first proposes 2,431 SMOTE- style interpolation candidates and retains 1,214 after rule filtering. State-dependent structural shocks generate 4,825 proposed minority-class candidates from the training- window distress observations and retain

2,431 after baseline and post-shock admissibility checks. The reported Δc values in Table 6 are the winsorized train -window mean differences between bankrupt and alive observations for each feature; complementary robustness, bootstrap, and decision-based summaries are provided in Tables 7 and Table 8. Additional ablation and SHAP analyses indicate that the cash-flow/liquidity block contributes most strongly, followed by profitability.